

2. Sze FK, Wong E, Or KK, Lau J, Woo J. Does acupuncture improve motor recovery after stroke? A meta-analysis of randomized controlled trials. *Stroke* 2002;33:2604-19.
3. Park J, Hopwood V, White AR, Ernst E. Effectiveness of acupuncture for stroke: a systematic review. *J Neurol* 2001;248:558-63.
4. Schnyer RN, Allen JJ. Bridging the gap in complementary and alternative medicine research: manualization as a means of promoting standardization and flexibility of treatment in clinical trials of acupuncture. *J Altern Complement Med* 2002;8:623-34.
5. Lewith GT, White PJ, Pariente J. Investigating acupuncture using brain imaging techniques: The current state of play. *Evid Based Complement Alternat Med* 2005;2:315-9.
6. Lee JD, Chon JS, Jeong HK, et al. The cerebrovascular response to traditional acupuncture after stroke. *Neuroradiology* 2003;45:780-4.
7. Lo YL, Cui SL, Fook-Chong S. The effect of acupuncture on motor cortex excitability and plasticity. *Neurosci Lett* 2005;384:145-9.
8. Jeun SS, Kim JS, Kim BS, et al. Acupuncture stimulation for motor cortex activities: a 3T fMRI study. *Am J Chin Med* 2005;33:573-8.

doi:10.1016/j.apmr.2006.02.011

The Post-Stroke Rehabilitation Outcomes Project Revisited

In December 2005, *Archives* published a supplement on the Post-Stroke Rehabilitation Outcomes Project (PSROP).¹ This project was a prospective observational cohort study that examined rehabilitation outcomes among 1291 consecutively enrolled stroke rehabilitation patients from 7 rehabilitation centers. A strength of the study was its detailed and clinically relevant characterization of the stroke rehabilitation process and how it related to outcomes at discharge. One finding was that stroke patients who received earlier and more aggressive therapy did better on certain dimensions than those who received less aggressive therapy. We recommended that these findings be further validated through predictive validity studies and possibly controlled trials.

Two commentaries^{2,3} on the PSROP spoke to its strengths and limitations. As study investigators, we want to respond to the comments because they address fundamental issues of rehabilitation research design and epistemology.

Jette² and Ottenbacher³ organize their commentaries around 4 sets of issues: selection bias, observation bias, confounds, and inference or interpretation. We use these issues to frame our discussion, and add a fifth, the relevance of evidence hierarchies.

Selection Bias

"Selection bias," says Ottenbacher, "occurs when there is a preferential inclusion of subjects with certain treatment outcomes."^{3(pS121)} Researchers can guard against selection bias in several ways. We guarded against it by using data on all consecutively admitted patients. This minimized, if not eliminated, selection bias due to preferential inclusion. Moreover, the study was not plagued by selection bias due to refusals during informed consent process or to drop-outs because the nature of the study did not require an informed consent process. Jette and Ottenbacher are correct that, at the subgroup level, there may have been subtle patient differences unknown to us and not captured in the statistical analyses. The exhaustiveness of the patient descriptors used in the study, however, helped to

minimize such differences. Even in randomized controlled trials (RCTs), there is the possibility of patient differences at the subgroup level if randomization is used for the entire study group but not at the subgroup level.

Observation Bias

Both Jette and Ottenbacher note that a "strength of the PSROP is the involvement of front-line clinicians in the development of the data collection instruments and actual data gathering and recording."^{3(pS121)} Such involvement ensures clinical relevance, but the downside is that the clinicians' awareness of the study's objectives may affect how they consciously or unconsciously treat patients and document patient outcomes.

In the PSROP, recording of interventions and documenting outcomes were done by different clinicians. Moreover, some outcomes are not as prone to observer bias as others. For example, recording discharge destination is not as prone to rater bias as recording a FIM score. It should be noted that there were no interventions pitted against another that would lead a therapist to favor 1 intervention over another. If therapists were prone to put their best foot forward, then the bias would be systematic across all interventions. Therapists did not know the outcomes or how the data would be analyzed. Our assessment is that after a brief self-conscious learning curve, data collection became routine and thus diminished the risk of observational bias.

One way to diminish observational bias is to blind treating therapists to the intervention. Blinding the treating clinician is easy in a drug study using a placebo, but it is impossible with most hands-on rehabilitation interventions. This is why some "more-pure" research designs do not work well in rehabilitation research.

Jette and Ottenbacher state that there is a lack of information about the accuracy or consistency of data in, and abstracted from, the medical chart. The supplement's methods article, by Gassaway et al,⁴ describes both the medical record abstraction process and the abstraction reliability process. Each medical records abstractor underwent a 4-day training session and a reliability testing process to ensure complete and accurate data collection. The study team checked for reliability at 4 points throughout the study and established a 95% agreement level between each chart abstractor and the study's lead trainer and reliability checker. Also, inaccurate data collection is likely to be more noisy and hence would bias against finding significant predictors.

Potential Confounders

An abiding concern about observational studies is that the relation between an intervention and an outcome may be confounded by other variables, one or more of which might not have been considered by the investigators. Researchers usually control for confounders through research design or statistical analyses, as in the case of the PSROP, that allowed researchers to control statistically for potential confounders. A strength of the PSROP, noted by the commentators, is that it captures a large number of potential confounders. Moreover, PSROP researchers relied heavily on the input of the study's clinical practice team to identify potential confounders.

One way the PSROP controlled for potential confounders was to use the Comprehensive Severity Index (CSI), an age- and disease-specific physiologic measure that encompasses more than 2000 signs, symptoms, and physical findings. The

commentators did not think that the CSI was adequately characterized in the supplement. But the variables for stroke diagnoses in the CSI are listed in appendix 4 of Gassaway⁴ and 8 references to the literature were given. It is difficult to display all variables in the CSI for common comorbidities associated with stroke because there can be hundreds. The validity and reliability of the CSI are discussed extensively in Gassaway.⁴

Ottenbacher notes correctly that the CSI did not enter several of the regression equations as a predictor variable and did not always displace the admission FIM score as a predictor of outcome. We sometimes found that arbitrarily entering 1 or the other of these 2 variables into a regression equation squeezed out the other as an explanatory variable. This finding corroborates Ottenbacher's observation that the 2 variables may share substantial variance. Interestingly, the investigators made a similar observation nearly 2 decades ago when the precursor to the FIM (the Barthel Index) and the precursor to the CSI (the Severity of Illness Index) were used in a stroke and hip fracture study.^{5,6} In the wake of the PSROP findings, we have proposed a study to separate criteria related to medical acuity from those related to function in the CSI to determine how each might independently affect outcome and to determine which medical acuity and function criteria are inextricably linked. More than anything, the PSROP findings speak to the robustness of the FIM, particularly the motor FIM, as a patient adjuster in stroke rehabilitation.

Regression analysis is a powerful statistical tool to control for confounders. Over 100 independent variables were considered. One concern raised by Jette and Ottenbacher is overspecification (ie, when the regression model has too many independent variables relative to the size of the study group or subgroup). While the overall study group was very large, commentators were rightfully concerned that analysis at the subgroup level could be overspecified thus increasing the risk of type I error. We did not overspecify the regression models. For the study group as a whole we used many independent variables but limited the number when doing subgroup analyses. We tested for interactions and took into account the assumptions required for complex regression analyses.

Inference and Interpretation

Ottenbacher argues that many statements in the supplement suggest causality when only association between variables can be inferred. He argues that the inconsistent use of terms implying both causation and association leads to proposing recommendations for change in practice that may be premature. We noted specifically in the opening article⁷ and elsewhere⁸ that association is not causality. When similar findings keep reoccurring, however, even after trying to explain them away with added confounders, one develops increasing confidence that the findings are not mere artifacts of research design or statistical analyses.

We proposed, moreover, that predictive validity studies be undertaken by actually implementing the findings and determining whether the same results occur. We also proposed in the supplement that 1 or more formal clinical trials be considered to validate the study's findings. The boundary between association and causation is sometimes thin; the path from association to causation should perhaps be construed as a continuum.

While the findings are compelling, PSROP investigators were appropriately cautious about study limitations. Ottenbacher states that only 1 limitation was mentioned in the introductory article,

but 5 are noted there and additional limitations are noted elsewhere.

Evidence Hierarchies

Ottenbacher proposes that the PSROP be classified as a level 3 or 4 study using criteria from the Center for Evidence-Based Medicine although an argument could be made that the PSROP is an example of a level 2c study using the same criteria.⁹ If Ottenbacher's classification is correct, it says as much about the limitations of the evidence hierarchy as it does about the strengths or limitations of the PSROP. The proposed classification also underscores the challenge that conventional evidence hierarchies present for rehabilitation and rehabilitation research. At the apex of this hierarchy is the RCT, and observational studies, regardless of how well designed, remain a distant third or fourth in the hierarchy. This hierarchy downgrades all well-designed observational studies like the PSROP because they fall short of the internal validity of RCTs, even though a small RCT with very strict patient selection criteria may be sorely lacking in external validity or generalizability. To suggest that a very small RCT—as many are—with narrow selection criteria is a stronger study than a very large observational study with broad selection criteria is to do a disservice to both clinical and health services research and to patients.

Invoking conventional evidence hierarchies as the standard also raises fundamental issues about the future of rehabilitation research. Without the possibility of multiparty blinding essential to randomized studies, should rehabilitation researchers just close up shop and move on to other fields that are amenable to traditional RCTs? Or should we develop research design alternatives to advance the field? Berwick challenges health services researchers to enrich the portfolio of methods to discover more effective systems of care. He states, “[h]ealth services research has not yet been sufficiently helpful in meeting the challenge of improving care in part because it has over-constrained both its methods and its favorite topics. The cost of insisting on formal, classical, summative, evaluative experimental designs in an uncertain, poorly understood, nonlinear, system is, unfortunately, to maintain the *status quo*.”^{10(p317)}

Ottenbacher observes that, in the case of hormone replacement therapy, “practice guidelines changed dramatically when large randomized trials did not support the findings of earlier observational studies.”^{3(pS122)} A counterexample from the supplement articles is worth noting.¹¹ RCTs sometimes report results that are not broadly applicable to real-world practice because of stringent selection criteria. In 1999, spironolactone was shown in a landmark RCT to reduce significantly deaths and hospitalizations among patients with congestive heart failure (CHF). In the 18 months that followed, there was a 4-fold increase in the number of prescriptions for spironolactone, followed by a tripling of hospital admissions and deaths resulting from dangerous elevations of potassium among CHF patients.¹² The problem was that many CHF patients who were prescribed spironolactone would have been excluded from the 1999 study. In short, the trial's selection criteria did not adequately take into account real-world practice. Trials tend to enroll the types of patients that maximize the chance of showing a benefit and minimize the chance of showing side effects. A complementary observational study could have provided the evidence needed to determine which types of patients could have benefited (or been harmed) in real-world practice.

Well-designed observational studies can provide timely discoveries that benefit both patients and clinicians. A classic example is the Framingham Heart Study that revolutionized our thinking about cardiovascular health and the role of life style interventions. Observational studies should not be down-

graded solely on philosophical grounds. We need practice-based evidence as much as we need evidence-based practice.

Gerben DeJong, PhD
National Rehabilitation Hospital
Washington, DC

Susan D. Horn, PhD
Randall J. Smout, MS
Julie Gassaway, MS, RN
Roberta James, Mstat
Institute for Clinical Outcomes Research
International Severity Information Systems Inc.
Salt Lake City, UT

References

- DeJong G, Horn SD, Gassaway J, Conroy B, editors. The Post-Stroke Rehabilitation Outcomes Project. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S1-125.
- Jette AM. The Post-Stroke Rehabilitation Outcomes Project. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S124-5.
- Ottensbacher KJ. The Post-Stroke Rehabilitation Outcomes Project. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S121-3.
- Gassaway J, Horn SD, DeJong G, Smout R, Clark C, James R. Applying the clinical practice improvement approach to stroke rehabilitation: methods used and baseline results. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S16-33.
- McGinnis GE, Osberg JS, DeJong G, Seward MS, Branch LG. Predicting charges for inpatient medical rehabilitation services using severity, DRG, age, and function. Am J Public Health 1987;77:826-9.
- Osberg JS, DeJong G, Haley S, Seward M, McGinnis G, Germaine J. Predicting long-term outcome among post-rehabilitation stroke patients. Am J Phys Med Rehabil 1988;67:94-103.
- DeJong G, Horn SD, Conroy B, Nichols D, Heaton EB. Opening the black box of poststroke rehabilitation: stroke rehabilitation patients, processes, and outcomes. Arch Phys Med Rehabil 2005; 86(12 Suppl 2):S1-7.
- Horn SD, DeJong G, Smout R, Gassaway J, James R, Conroy B. Stroke rehabilitation patients, practice, and outcomes: is earlier and more aggressive therapy better? Arch Phys Med Rehabil 2005;86(12 Suppl 2):S101-14.
- Center for Evidence-Based Medicine. Levels of evidence and grades of recommendation. Available at: http://www.cebm.net/levels_of_evidence.asp#notes. Accessed February 2, 2006.
- Berwick DM. The John Eisenberg lecture: health services research as a citizen in improvement. Health Serv Res 2005;40:317-36.
- Horn SD, DeJong G, Ryser DK, Veazie PJ, Teraoka. Another look at observational studies in rehabilitation research: going beyond the holy grail of the randomized controlled trial. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S8-15.
- McMurray JJ, O'Meara E. Treatment of heart failure with spironolactone—trail and tribulations. N Engl J Med 2004;351:526-8.

doi:10.1016/j.apmr.2006.02.009

The author replies

I must confess I am more than a bit perplexed by DeJong et al's spirited defense in their letter because I am in total agreement with their statement that "Observational studies should not be downgraded solely on philosophical grounds," as even a

cursory reading of my commentary¹ would reveal. As I stated in my commentary,

The PSROP [Post Stroke Rehabilitation Outcomes Project] provides an important example of the value of observational study designs in rehabilitation, and I applaud the investigators for their important accomplishment, one that I hope is replicated by others. The PSROP provides us with an important additional method to respond to calls for the rehabilitation field to demonstrate the effectiveness of the services it provides.^{1(pS125)}

To support a methodology is not the same as ignoring its limitations. All research designs have limitations and, in my commentary on the PSROP, I attempted to summarize what I saw as some of the study's major strengths as well as some of its important limitations, particularly concerns with internal validity and generalizing the findings from 1 observational study to clinical practice. I stand by those concerns.

For me, DeJong's argument against "The cost of insisting on formal, classical, summative, evaluative experimental designs in an uncertain, poorly understood, nonlinear, system is, unfortunately, to maintain the *status quo*"^{2(p317)} is based on a false and unhelpful dichotomy. The choice is not randomized controlled trial designs versus observational designs, but the appropriate application of both to help advance rehabilitation science, to move rehabilitation firmly onto an evidence-based foundation.

Both well-designed observational and experimental studies can and will provide timely discoveries that benefit patients and clinicians. As a researcher who has applied both observational and experimental designs in my own research, it is obvious to me that both are valuable and need to be applied with increasing frequency in the rehabilitation field.

Alan M. Jette, PT, PhD
Health & Disability Research Institute
Boston University
Boston, MA

References

- Jette AM. The Post-Stroke Rehabilitation Outcomes Project. Arch Phys Med Rehabil 2005;86(12 Suppl 2):S124-5.
- Berwick DM. The John Eisenberg lecture: health services research as a citizen in improvement. Health Serv Res 2005;40:317-36.

doi:10.1016/j.apmr.2006.02.010

The author replies

I appreciate the response by DeJong and others to my commentary on the Post-Stroke Rehabilitation Outcomes Project (PSROP) articles. I concur with many of their comments, but was surprised by the defensive nature of the remarks regarding evidence hierarchies. I agree that rehabilitation researchers should use a variety of approaches to establish evidence-based rehabilitation and stated this in my commentary.¹ Regarding the study limitations, DeJong et al² note that I list only 1 limitation from the first article in the supplement.³ The manuscript I reviewed in drafting my commentary included the following statement under a heading titled "PSROP's Chief Limitation":